

Attorney Docket No.: 019680-008500US
Client Reference No.: P001163

PATENT APPLICATION

DEADLOCK AVOIDANCE IN A BUS FABRIC

Inventor: David G. Reed, a citizen of Canada, residing at
18801 Ten Acres Road
Saratoga, CA 95070

Assignee: NVIDIA Corporation
2701 San Tomas Expressway
Santa Clara, CA 95050

Entity: Large

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, Eighth Floor
San Francisco, California 94111-3834
Tel: 650-326-2400

DEADLOCK AVOIDANCE IN A BUS FABRIC

BACKGROUND

[0001] The present invention relates generally to deadlock avoidance in a bus fabric, and more particularly to deadlock avoidance at an interface between integrated circuits.

[0002] Few applications stress the resources of a computer systems to the extent that video does. Video capture, encoding, and the like involve huge transfers of data between various circuits in a computer system, for example, between video-capture cards, central processing units, graphics processors, systems memories, and other circuits.

10 [0003] Typically, this data is moved over various buses, such as PCI buses, HyperTransport™ buses, and the like, both on and between the integrated circuits that form the computer system. Often, first-in-first-out memories (FIFOs) are used to isolate these circuits from one another, and to reduce the timing constraints of data transfers between them.

[0004] But these FIFOs consume expensive integrated circuit die area and power.

15 [0005] Accordingly, it is desirable to limit the depth of the FIFOs. Unfortunately, this means that these FIFOs may become filled and not able to accept further inputs, thus limiting system performance.

[0006] It is particularly problematic if these filled FIFOs are in a data path that forms a loop. In that case, there may be a processor, such as a graphics processor, or other circuit in the loop 20 that becomes deadlocked, that is, unable to either receive or transmit data.

[0007] This can happen under the following conditions, for example. A first FIFO that receives data from a circuit cannot receive data because it is full. The first FIFO cannot send data to a second FIFO because the second FIFO is also full. The second FIFO similarly cannot send data because it wants to send the data to the circuit, which cannot accept it since it is 25 waiting to send data to the first FIFO. This unfortunate set of circumstances can result in a stable, deadlocked condition.

[0008] Thus, what is needed are circuits, methods, and apparatus for avoiding these deadlocked conditions. While it may alleviate some deadlocked conditions to increase the size of the FIFOs, again there is an associated cost in terms of die area and power, and the possibility

remains that an even deeper FIFO may fill. Thus, it is desirable that these circuits, methods, and apparatus not rely solely on making these FIFOs deeper and be of limited complexity.

SUMMARY

5 [0008] Accordingly, embodiments of the present invention provide circuits, apparatus, and methods for avoiding deadlock conditions. One exemplary embodiment provides an address decoder for determining whether a received posted write request is a peer-to-peer request. If it is, the request is converted to a non-posted write request. A limit on the number of pending non-posted requests is maintained and not exceeded, such that deadlock is avoided. The number of
10 pending non-posted requests is tracked by subtracting the number of responses received from the number of non-posted requests sent.

[0009] Another exemplary embodiment does not convert received posted requests to non-posted requests, but rather provides an arbiter that tracks the number of pending posted requests. When the number of pending posted requests (for example, the number of pending
15 requests in a FIFO or queue) reaches a predetermined or programmable level, that is a low-water mark, a Block Peer-to-Peer signal is sent to an arbiter's clients. This keeps the FIFOs in a data loop from filling, thus avoiding deadlock. When a response or signal indicating that the number of pending posted requests is below this level is received by the arbiter, the Block Peer-to-Peer signal is removed, and peer-to-peer requests may again be granted. Alternately, the number of
20 pending peer-to-peer requests may be tracked, and when a predetermined or programmable level is reached, a Block Peer-to-Peer signal is asserted. Circuits, methods, and apparatus consistent with the present invention may incorporate one or both of these or the other embodiments described herein.

[0010] A further exemplary embodiment of the present invention provides a method of
25 transferring data. This method includes receiving a transfer request, determining if the transfer request is a write to a memory location, if the transfer request is a write to a memory location, then sending the transfer request as a posted request, otherwise determining a number of available transfer request entries in a posted-request first-in-first-out memory, and if the number of transfer request entries available is greater than a first number, then sending the transfer request as a posted request, otherwise waiting to send the transfer request as a posted request.
30

[0011] A further exemplary embodiment of the present invention provides another method of transferring data. This method includes maintaining a first number of tokens, receiving a plurality of posted requests, if a remaining number of the first number of tokens is less than a first number, forwarding one of the plurality of posted requests as a non-posted request, else not 5 forwarding the one of the plurality of posted requests as a non-posted request.

[0012] Yet another exemplary embodiment of the present invention provides An integrated circuit. This integrated circuit includes an arbiter configured to track a number of available entries in a posted request FIFO, a plurality of clients coupled to the arbiter, and a HyperTransport bus coupled to the arbiter, wherein the arbiter receives peer-to-peer requests 10 from the plurality of clients and provides posted requests to the posted request FIFO, and when the number of available entries in the posted request FIFO is equal to a first number, then preventing the plurality of clients from sending peer-to-peer requests.

[0013] A better understanding of the nature and advantages of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

15

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] Figure 1 is a block diagram of a computing system that benefits by incorporation of embodiments of the present invention;

[0015] Figure 2 is a block diagram of an improved computing system that is benefited by the 20 incorporation of embodiments of the present invention;

[0016] Figure 3 is a simplified block diagram of the improved computing processing system of Figure 2;

[0017] Figure 4 is a further simplified block diagram of the improved computing system of Figure 2 illustrating the write path from a video-capture card to a system memory;

[0018] Figure 5 is a simplified block diagram of the improved computing system of Figure 2 that incorporates an embodiment of the present invention;

[0019] Figure 6 is a flowchart further describing a specific embodiment of the present invention;

[0020] Figure 7 is a simplified block diagram of the improved computing system of Figure 2 that incorporates an embodiment of the present invention; and

[0021] Figure 8 is a flowchart further describing a specific embodiment of the present invention.

DESCRIPTION OF EXEMPLARY EMBODIMENTS

5 [0022] Figure 1 is a block diagram of a computing system 100 that benefits by incorporation of embodiments of the present invention. This computing system 100 includes a Northbridge 110, graphics accelerator 120, Southbridge 130, frame buffer 140, central processing unit (CPU) 150, audio card 160, Ethernet card 162, modem 164, USB card 166, graphics card 168, PCI slots 170, and memories 105. This figure, as with all the included figures, is shown for illustrative
10 purposes only, and does not limit either the possible embodiments of the present invention or the claims.

[0023] The Northbridge 110 passes information from the CPU 150 to and from the memories 105, graphics accelerator 120, and Southbridge 130. Southbridge 130 interfaces to external communication systems through connections such as the universal serial bus (USB) card 166 and
15 Ethernet card 162. The graphics accelerator 120 receives graphics information over the accelerated graphics port (AGP) bus 125 through the Northbridge 110 from CPU 150 and directly from memory or frame buffer 140. The graphics accelerator 120 interfaces with the frame buffer 140. Frame buffer 140 may include a display buffer that stores pixels to be displayed.

20 [0024] In this architecture, CPU 150 performs the bulk of the processing tasks required by this computing system. In particular, the graphics accelerator 120 relies on the CPU 150 to set up calculations and compute geometry values. Also, the audio or sound card 160 relies on the CPU 150 to process audio data, positional computations, and various effects, such as chorus, reverb, obstruction, occlusion, and the like, all simultaneously. Moreover, the CPU 150 remains
25 responsible for other instructions related to applications that may be running, as well as for the control of the various peripheral devices connected to the Southbridge 130.

[0025] Figure 2 is a block diagram of an improved computing system that is benefited by the incorporation of embodiments of the present invention. This block diagram includes a combined processor and Northbridge 210, media control processor 240, and system memory 270. Also
30 included in this block diagram for exemplary purposes is a video capture card 280.

[0026] The combined processor and Northbridge 210 includes a central processing unit 212, FIFO 216, multiplexer 222, output buffers 224 including one for posted requests 226, non-posted requests 228, and responses 230, input FIFO 232 including an input FIFO for posted requests 234, non-posted requests 236, and responses 238, address decoder 220, peer-to-peer FIFO 218, and memory controller 214.

[0027] The media control processor includes input FIFO 242 for posted requests 244, non-posted requests 246, and responses 248, an integrated graphics processor 252, arbiter 250, and PCI-to-PCI bridge 260. The combined CPU and Northbridge 210 communicates with the media control processor 240 over HyperTransport buses 290 and 295. The system memory 270 couples to the memory controller 214 over memory interface bus 272, while the video capture card 280 is connected to the PCI-to-PCI bridge 260 over the PCI bus 282.

[0028] In a specific embodiment of the present invention, the combined CPU and Northbridge 210 is formed on a first integrated circuit, while the media control processor 240 is formed on a second integrated circuit. In another embodiment, the graphics processor 252 is not integrated on the media control processor, but is rather a separate integrated circuit communicating over an advanced graphics processor (AGP) bus with the media control processor 240. In other embodiments, these various functions may be divided in other ways and integrated on different numbers of integrated circuits communicating over various buses.

[0029] Data and requests move between these integrated circuits and integrated circuit blocks over buses. In the case of a write request, a circuit requests that it be allowed to place data on a bus, and that request is granted. The data may either be sent as a posted request, in which no response is required, or as a non-posted request, in which case a response is required. The response is sent back to the sending circuit after the write has been completed at its destination circuit.

[0030] These different transactions, posted requests, non-posted requests, and responses, are stored in separate FIFOs as shown. These separate FIFOs may be the same size, or they may be different sizes. Further, the various FIFOs may have different sizes. In one specific embodiment, the non-posted request FIFO 236 has 6 entries, the peer-to-peer FIFO 218 has two entries, and the non-posted request FIFO 228 has 16 entries. In various embodiments, the peer-to-peer FIFO 218 may be one FIFO for storing posted and non-posted requests and responses, or it may be separate FIFOs for storing the different types of transactions. More information about

these various types of requests and peer-to-peer transactions can be found in the HyperTransport specification, which is currently on release 1.05 published by the HyperTransport Consortium, which is incorporated by reference.

[0031] In this new architecture, the graphics processor has become separated from the system memory. This separation leads to data paths that can form a loop, and thus become deadlocked. Specifically, data transfers from the CPU 212 and video capture card 280 may fill the various FIFOs.

[0032] In the configuration shown in Figure 2, the CPU 212 writes to a frame buffer in the system memory 270 utilizing the following path. The CPU 212 provides requests (data), on line 10 213 to the FIFO 216. The FIFO 216 provides data to the multiplexer 222, which in turn provides the data to the output buffers 224. The buffers 224 provide data over HyperTransport bus 290 to FIFO 242, which in turn provide data to the graphics buses 252. The graphics processor 252 provides the requests on line 254 to the arbiter 250. The arbiter 250 provides the requests back over the HyperTransport bus 295 to the FIFO 232. The FIFO 232 provides the request to the 15 address decoder 220, which in turn provides them to the memory controller 214. The memory controller 214 writes to the system memory 270 over memory interface buses 272.

[0033] Also in this configuration, the video capture card 218 writes data to a frame buffer in the systems memory 270 utilizing the following path. The video capture card 280 provides data on PCI bus 282 to the PCI-to-PCI bridge 260. The PCI-to-PCI bridge 260 provides data to the 20 arbiter 250, which in turn provides the requests over HyperTransport bus 295 to the FIFO 232. The FIFO 232 provides the requests to the address decoder 220, which in turn provides it to the peer-to-peer FIFO 218. The peer-to-peer FIFO 218 provides the data to multiplexer 222, which in turn provides it to the output buffers 224. The output buffers 224 provide the data to the FIFO 242, which in turn provides it to the graphics processor 252.

[0034] The graphics processor 252 then writes to the frame buffer in the systems memory 270 utilizing the following path. The graphics processor 252 provides modified requests on line 254 to the arbiter 250. The arbiter 250 provides the data over HyperTransport bus 295 to the FIFO 232. FIFO 232 provides the data to the address decoder 220. This time, the address decoder sees a new address provided by the graphics processor 252, and in turn provides the request to 30 the memory controller 214. The memory controller 214 then writes the data to the systems memory 270 over memory interface buses 272.

[0035] As can be seen, this convoluted path crosses the HyperTransport interface buses 290 and 295 a total of three times. Particularly in situations where the CPU 212 and video capture card 280 are writing to a frame buffer in the systems memory 270, the FIFOs 242, 232, and 218 may become full, that is, unable to accept further inputs. In this case, the situation may arise
5 where the graphics processor 252 tries to write data to the frame buffer in the systems memory 270, but cannot because the arbiter 250 can not grant the graphics processor 252 access to the HyperTransport bus 295. Similarly, the receive FIFO 232 cannot output data because the peer-to-peer FIFO 218 is full. Further, the peer-to-peer FIFO 218 cannot output data because the media control processor input FIFO 242 is similarly full. In this situation, the bus fabric is
10 deadlocked and an undesirable steady-state is reached.

[0036] Figure 3 is a simplified block diagram of the improved computing processing system of Figure 2. Included are a combined CPU and Northbridge 310, media control processor 340, system memory 370, and video capture card 380. The combined CPU and Northbridge 310 includes a transmitter 312 and receiver 314, while the media control processor includes a
15 receiver 342, transmitter 344, graphics processor 346, and PCI-to-PCI bridge 348. A systems memory 370 communicates with the combined CPU and Northbridge over a memory interface bus 372. In this particular example, a video capture card 380 is included, which communicates with the media control processor over a PCI bus 382.

[0037] Figure 4 is a further simplified block diagram of the improved computing system of Figure 2 illustrating the write path from the video-capture card 480 to the system memory 470. This block diagram includes a combined CPU and Northbridge 410, media control processor 440, system memory 470, and video capture card 480. The combined CPU and Northbridge circuit includes a transmitter 412 and receiver 414. The media control processor includes a receiver 442, transmitter 444, graphics processor 446, and PCI-to-PCI bridge 448.
20

[0038] The video capture card 480 provides requests to the PCI-to-PCI bridge 448, which in turn provides them to the transmitter 444. The transmitter 444 sends requests to the receiver 414, which in turn provides them to the transmitter 412. The transmitter 412 sends these requests to the receiver 442, which passes them along to the graphics processor 446. The graphics processor 446 writes the data to the systems memory by sending it as a request to the transmitter 444, which in turn provides it to the receiver 414. The receiver 414 then writes the data to the
25 systems memory 470.
30

[0039] As can be seen, the requests cross from the transmitter 444 to the receiver 414 twice during this process. This is where the potential for a deadlock arises. Specifically, in the deadlocked condition, the graphics processor cannot send a request to the transmitter 444, because the transmitter cannot send to the receiver 414, since its associated FIFO is full. The 5 graphics processor cannot accept a new request because it is waiting to granted its own request. Accordingly, it cannot drain the FIFO in the receiver 442. Again, a deadlocked condition arises, creating an undesirable steady-state.

[0040] Figure 5 is a simplified block diagram of the improved computing system of Figure 2 that incorporates an embodiment of the present invention. This block diagram includes a 10 combined CPU and Northbridge 510, media control processor 540, systems memory 570, and video card 580. The combined CPU and Northbridge 510 includes a transmitter 512 and a receiver 514. The media control processor 540 includes a receiver 542, transmitter 544, graphics processor 546, and PCI-to-PCI bridge 548. The PCI-to-PCI bridge 548 further includes an address decoder 562.

[0041] A posted request is provided by the video capture card 580 to the PCI-to-PCI bridge 548. The address decoder 562 in the PCI-to-PCI bridge 548 determines that this posted request is a peer-to-peer request and converts it to a non-posted request and passes it to the transmitter 544. The transmitter 544 sends this request as a non-posted request that is sent to the receiver 514. The receiver 514 then sends the request to the transmitter 512, which passes it to the 20 receiver 542. The receiver 542 in turn provides the request to the graphics processor 546.

[0042] The graphics processor 546 then reflects the request back upstream to the transmitter 544 as a posted request having an address in the frame buffer in the system memory 570. The graphics processor also issues a "target done" completion response. The combined CPU and Northbridge 510 receive the posted request and response from the transmitter 544. The posted 25 request is sent to the system memory 570, and the response is sent back to the media control processor 540, where it is received by the PCI-to-PCI bridge 548.

[0043] In this embodiment, the number of pending non-posted requests is limited to some number "N", such as 1, and when this number is reached, no further non-posted requests are provided to the transmitter 544. Specifically, as a non-posted request is sent, a count is 30 incremented in the address decoder portion 562 of the PCI-to-PCI bridge 548. As responses are received by the PCI-to-PCI bridge 548, this count is decremented. When the count is reached,

further non-posted requests are held by the address decoder 562. This avoids the deadlocked condition described above.

[0044] Figure 6 is a flowchart further describing this specific embodiment of the present invention. In act 610, a posted request is received from a video capture card. In act 620, the address associated with the request is decoded and a determination of whether the request is peer-to-peer or to be written to the system memory is made. If it is not a peer-to-peer request, that is, it is data to be written to the system memory, it is sent as a posted request in act 680. If it is a peer-to-peer request, the request is converted to a non-posted request in act 630. In act 640, it is determined whether the number of pending non-posted requests is equal to a predetermined or programmable number of allowable pending non-posted requests, such as 1 or another number, in act 650. If the count has not reached this number "N", the request may be sent as a non-posted request in act 660, and the count is incremented by one in act 670. If the count has reached "N" however, the requests is stalled or not granted in order to avoid a deadlocked condition in act 650. As non-posted requests are completed, responses are received and the count is decremented.

[0045] Returning to Figure 2, we can see how this embodiment is implemented in greater detail. The video capture card 280 provides posted requests on PCI bus 282 to the PCI-to-PCI bridge 260. An address decoder in the PCI-to-PCI bridge 260 determines whether the request is a write to the system memory 270. If it is, that request is passed to the arbiter 250 which places it in the posted request FIFO 234, which forwards it to the memory controller 214, which writes it to the system memory 270.

[0046] If the posted request is a peer-to-peer request, that is, it is not to be written directly to the system memory 270 but is destined for a peer circuit, for example the graphics processor 252, then the posted request is converted to a non-posted request by an address decoder (or other circuit) in the PCI-to-PCI bridge 260. This non-posted request is routed from the arbiter 250 to the non-posted request FIFO 236, to the peer-to-peer FIFO 218. The non-posted request then reaches the graphics processor 252 via bus 290. The graphics processor converts the request back to a posted request and also issues a response. The posted request is passed to the memory controller 214 which writes the data to the system memory 270, while the response is received by the PCI-to-PCI bridge 260.

[0047] The decoder in the PCI-to-PCI bridge 260 also keeps track of the number of pending non-posted requests, and does not send non-posted requests to the non-posted request FIFO 236 once it has determined that a predetermined or programmable number of pending non-posted requests has been reached.

5 [0048] Figure 7 is a simplified block diagram of the improved computing system of Figure 2 that incorporates an embodiment of the present invention. This block diagram includes a combined CPU and Northbridge 710, media control processor 740, systems memory 770, and video-capture card 780. The combined CPU and Northbridge 710 includes a transmitter 712 and receiver 714. The media control processor 740 includes a receiver 742, transmitter 744, graphics processor 746, and PCI-to-PCI bus 748. The transmitter 744 further includes an arbiter 745.

10 [0049] Posted requests provided by the video capture card 780 are provided to the PCI-to-PCI bridge 748, which passes them to the arbiter 745. The arbiter tracks posted requests (or alternately, peer-to-peer requests) that are pending at the receiver 714. When a certain number of posted requests remain pending, the arbiter 745 sends out a Block Peer-to-Peer signal to its

15 clients such as the graphics processor 746 and PCI-to-PCI bridge 748. In this case, no further peer-to-peer requests are sent to the arbiter 745 until a response indicating that there is room in the receiver 714 posted request FIFO is received by the arbiter 745.

[0050] If the Block Peer-to-Peer signal is not asserted, the posted request is provided to the transmitter 744, which sends it to the receiver 714. The receiver 714 routes it to the receiver 742 via the transmitter 712. The receiver 742 passes the posted request to the graphics processor 746. The graphics processor 746 in turn passes it to the transmitter 744 to the receiver 714, which provides it to the system memory 770.

20 [0051] Figure 8 is a flowchart further describing this specific embodiment of the present invention. In act 810, an arbiter receives a posted request, for example from a video capture card. In act 820, the arbiter determines whether the posted request is a peer-to-peer request. If it is not, then in act 830, the data is sent as a posted write request. If it is, then in act 840, it is determined whether the FIFO is below its low-water mark, or alternatively, whether a block peer-to-peer signal or state has been asserted. If this is true, then in act 850, the arbiter waits for an entry to become available in the posted write FIFO. At some point, the posted write FIFO provides an output, thus freeing up an entry. At this time, the arbiter releases the Block Peer-to-Peer signal and the data is sent to the posted write FIFO in act 830.

[0052] Returning to Figure 2, we can see how this embodiment is implemented in greater detail. The video capture card 280 provides posted requests on PCI bus 282 to the PCI-to-PCI bridge 260. The PCI-to-PCI bridge 260 passes these requests to the arbiter 250. The arbiter keeps track of a number of pending posted requests in the posted request FIFO 236 (or 5 alternately, the number of pending peer-to-peer requests, or the number of posted requests in FIFO 218). When the number of pending posted requests in the posted request FIFO 236 reaches a predetermined or programmable level the arbiter 250 broadcasts a Block Peer-to-Peer signal to the graphics processor 252, PCI-to-PCI bridge 260, and other client circuits. This keeps those circuits from sending further peer-to-peer requests, thus avoiding a deadlocked condition.

10 [0053] When the number of pending posted requests is below this low-water mark, the posted request is sent to the posted request FIFO 234. The posted request is then routed through the peer-to-peer FIFO 218, multiplexer 222, FIFOs 226 and 244, to the graphics processor 252. The graphics processor then converts the address to a system memory address 270, and forwards the posted request to the arbiter 250. The arbiter 250 passes the posted request to the posted request 15 FIFO 234, to the memory controller 214, which writes data to the system memory 270.

[0054] In one embodiment, at power up, the arbiter 250 receives a number of tokens, for example six tokens. As the arbiter provides a peer-to-peer posted request to the posted request FIFO 234, it sends along one of these tokens. As the posted request FIFO outputs a peer-to-peer posted request, the arbiter receives a token. If the count of tokens drops to a low-water mark 20 level, for example one, the arbiter 250 asserts the Block Peer-to-Peer signal. When tokens are received, the Block Peer-to-Peer signal is removed.

25 [0055] The above description of exemplary embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above. The embodiments were chosen and described in order to best explain the principles of the invention and its practical applications to thereby enable others skilled in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated.